

Weekly Report

Pingping Shang

2013.12.30~2014.1.5

本周工作

- 1、因为可能之后我的工作也加入到经济数据的可视分析上，所以整理淘宝数据的系统框架，保证选择、查询、筛选、重排、修正各个交互能正确响应，暂时将其搁置。
- 2、熟悉解聪编写的处理经济数据的系统框架，因为淘宝数据是纯 java 编写，经济数据采用 processing，很多地方不同，熟悉系统结构便于后面的参与。
- 3、对照我们的可视分析框架，总结离散随机变量可视化的对应模型和实际交互，具体总结见附页。
- 4、对照可视化关联规则的文章（Visualizing Association Rules with Interactive Mosaic Plots），跟我们的工作做对比，貌似我们现在实现的功能跟它的差不多，没有更多新意。后面有总结那篇文章的思路。

下周工作

参与到经济数据可视分析的架编写中，前面所说总结还是不太清晰，还会再和解聪讨论，总结明确了，系统就可按需求进行完善。

一、符号定义

A_i : 条件属性

B_i : 目标属性

A_N : 条件属性个数

B_N : 目标属性个数

$Block_{(x,y)}$: 第 x 行第 y 列的 block

N_{xi} : 竖直方向第 i 个属性的属性值数目

N_{yi} : 水平方向第 i 个属性的属性值数目

BX : Block 矩阵行数, $BX = \prod N_{xi}$

BY: Block 矩阵列数, $BY = \prod N_{yi}$

$Cell_{(x,y)}$: 第 x 行第 y 列的 cell

C_X: Block 矩阵行数, $C_X = \prod N_{xi}$

C_Y: Block 矩阵列数, $C_Y = \prod N_{yi}$

pdf: 条件属性取一组特定值时, 目标属性的概率分布

二、操作总结

交互	操作	模型
选取与标记	1. Block 单选: 鼠标点击 $Block_{(x,y)}$.	$Block_{(x,y)}$ 的条件属性值为 $S=\langle a_1, a_2 \dots a_{A_N} \rangle$, 条件属性 A_1 取值 a_1 , A_2 取值 $a_2 \dots A_{A_N}$ 取值 a_{A_N} 。其对应的目标属性的条件分布为 $P(b_1, b_2 \dots b_{B_N} a_1, a_2 \dots a_{A_N})$
	2. Block 圈选: 使用交互工具圈选 $Block_{(x,y)}$ 等一组 Block	条件属性取 $S_1, S_2 \dots S_N$ 等属性值组合对应的目标属性 B 的条件分布
	3. 整行整列: 鼠标点击 Block 矩阵某行 (列) 表头	一组条件属性取值固定, 剩余条件属性取各种可能值对应的目标属性实例集 X, 其中 $0 \leq i \leq A_N$
	4. Cell 划选: 使用交互工具圈选 $Cell_{(x,y)}$ 等一组 Cell.	条件属性取值 S, 目标属性 $B_1, B_2 \dots B_i$ 的概率分布律, 其中 $0 \leq i \leq B_N$ $P(b_1, b_2 \dots b_i a_1, a_2 \dots a_{A_N})$
查找	1. 使用 选取交互 2 选中 $Block_{(x,y)}$, 查找相似 Block	查找与 $Block_{(x,y)}$ 的距离小于 δ 的 Block $d(pdf_{(x,y)}, pdf) < \delta$ 其中 Block 距离定义: $d(pdf_1, pdf_2) = d_A(pdf_1, pdf_2)$ $= \ pdf_1 - pdf_2 \ _A$ $= \sqrt{(pdf_1 - pdf_2)^T A (pdf_1 - pdf_2)}$
	2. 先使用 选取交互 2 选中 $Block_{(x,y)}$, 再用 选取交互 4 选中目标属性取值范围.	处于取值范围内的目标属性权重为 1, 即距离公式 A 中对应元素为 1, 否则权重为 0, A 中对应元素为 0
	鼠标移动到行 (列) 表头, 点击删	一组条件属性取值固定, 剩余条件属性取各个可

筛选	除该行（列），该删除可恢复	能值时目标属性 B 的概率分布特征都不明显，将其删除
修正	使用 选取交互 3 ，选取一组 Block，将其作为相似项输入	<p>指定条件属性取值$S_1, S_2 \dots S_i$时，目标属性 B 的概率分布相似，以此作为约束条件学习距离定义中的 A</p> $\min_A \sum_{(pdf_i, pdf_j) \in S} \ pdf_i - pdf_j\ _A^2$ $s.t. \sum_{(pdf_i, pdf_j) \in D} \ pdf_i - pdf_j\ _A^2,$ $A \geq 0.$
重排	鼠标拖动矩阵行（列）表头，红线指示可释放位置，按需求对条件属性取值排序	<p>一组条件属性分别取值$S_1 = \langle a_1, a_2 \dots a_i \rangle$和$S_2 = \langle a'_1, a'_2 \dots a'_i \rangle$，剩余条件属性取值变化时，对应目标属性 B 的概率分布有比较意义，则将S_1, S_2对应的 B 的概率分布相邻排放</p>

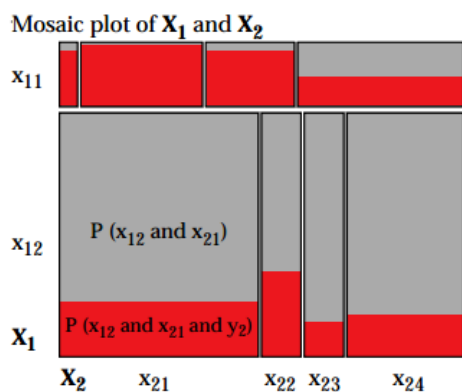
[2000] Visualizing Association Rules with Interactive Mosaic Plots

1 关联规则：X→Y

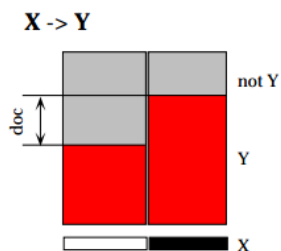
支持度：counts(X^Y) / counts(U) U 表示所有条目的集合

置信度：counts(X^Y) / counts(X)

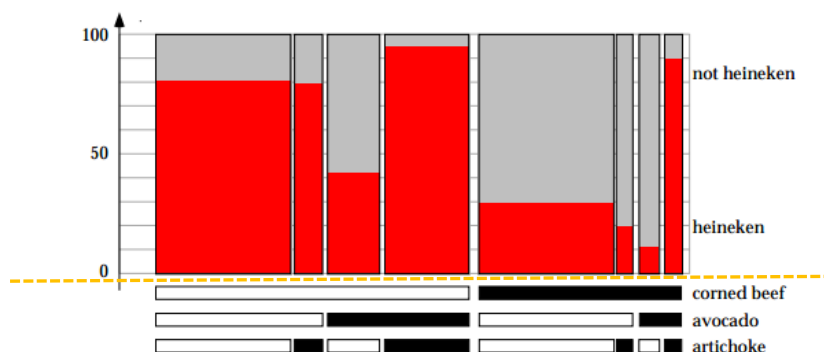
2 马赛克图中每块的面积表示支持度(相对大小)，也可以理解为联合概率，highlighted 区域占所在块的比例为置信度，如下图所示



3 当某条超过设定阈值的规则被挖掘出后，到底置信度比其他 rules 大多少，人们没有概念，马赛克图即可弥补这个缺点，如下图所示



4 Double Decker Plot ---只在一个方向划分，但是有另一面板来说明各属性的取值范围，如下图所示，橙色线以上为只在水平方向划分后的结果，橙色线以下每行代表一次划分，即增加一个属性，通过跟上面相邻行的比较，可得知该属性可取值的个数，条的宽度代表值大小。



这种方法可直观看到置信度的差异。

5 交叉结构

两条规则 $X_1 \rightarrow Y$ 和 $X_2 \rightarrow Y$ ，很有可能描述的是同一个人群。

举个例子，1) 买牛奶的人会买面包，2) 买奶油的人会买面包，规则 1) 2) 很可能描述的是同一群人，即买牛奶和买奶油的本来就是一群人，这样得到的信息就会片面，马赛克图通过交叉混合两条规则，可看出 X_1 和 X_2 是否是独立的，即买牛奶和奶油的人是不是同一人群。

6 序列结构关联规则的探索：

$$\begin{aligned} X_1 &\rightarrow Y \\ X_1 \wedge X_2 &\rightarrow Y \\ X_1 \wedge X_2 \wedge X_3 &\rightarrow Y \end{aligned}$$

序列结构如右图所示，通过增加条件属性得到另外的规则：

...

按添加属性的顺序展示各条规则对应的马赛克图，探索属性的增加对置信度和支持度的影响。